REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
25-02-2011	Final	Feb. 2008- Nov. 2010

4. TITLE AND SUBTITLE
Adaptive Information Filtering: Final Report

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-08-1-0074

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Yi Zhang, University of California Santa Cruz

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California Santa Cruz
SOE3, 1156 High Street, Santa Cruz, CA 95064

8. PERFORMING ORGANIZATION REPORT NUMBER
NA

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolfph St. Room 3112
Arlington VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-0215

12. DISTRIBUTION/AVAILABILITY STATEMENT
Public

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project studies personalized proactive information filtering agents that pushes relevant information to the user without requiring explicit user query. To do this, the agent adaptively learns a detailed user model while observing and interacting with the user. We use Bayesian statistical theory and machine learning techniques to tackle the following two major challenges. We studied two major problems: how to build an initial user profile with minimal user effort, and How to improve personalized recommendation based on multiple evidences, such as social networks, implicit user feedback, and explicit user feedback and context information. This project led to 1 book chapter, 2 journal paper, 4 conference publications and one demo system.

15. SUBJECT TERMS
information filtering, information retrieval, text processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Yi Zhang
U	U	U	UU		19b. TELEPHONE NUMBER *(Include area code)* 8314594549

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18
Adobe Professional 7.0

Adaptive Information Filtering
Abstract

Yi Zhang, University of California Santa Cruz

This project studies personalized proactive information filtering agents that pushes relevant information to the user without requiring explicit user query. To do this, the agent adaptively learns a detailed user model while observing and interacting with the user. We use Bayesian statistical theory and machine learning techniques to tackle the following two major challenges. We studied two major problems: how to build an initial user profile with minimal user effort, and How to improve personalized recommendation based on multiple evidences, such as social networks, implicit user feedback, and explicit user feedback and context information. The research has be evaluated on TREC data, data the PI has collected through a user study, and data collected from digg.com, Citeseer, del.io.us.. This project led to 1 book chapter, 2 journal paper, 4 refereed conference publications, 1 un-refereed publication and one demo system.

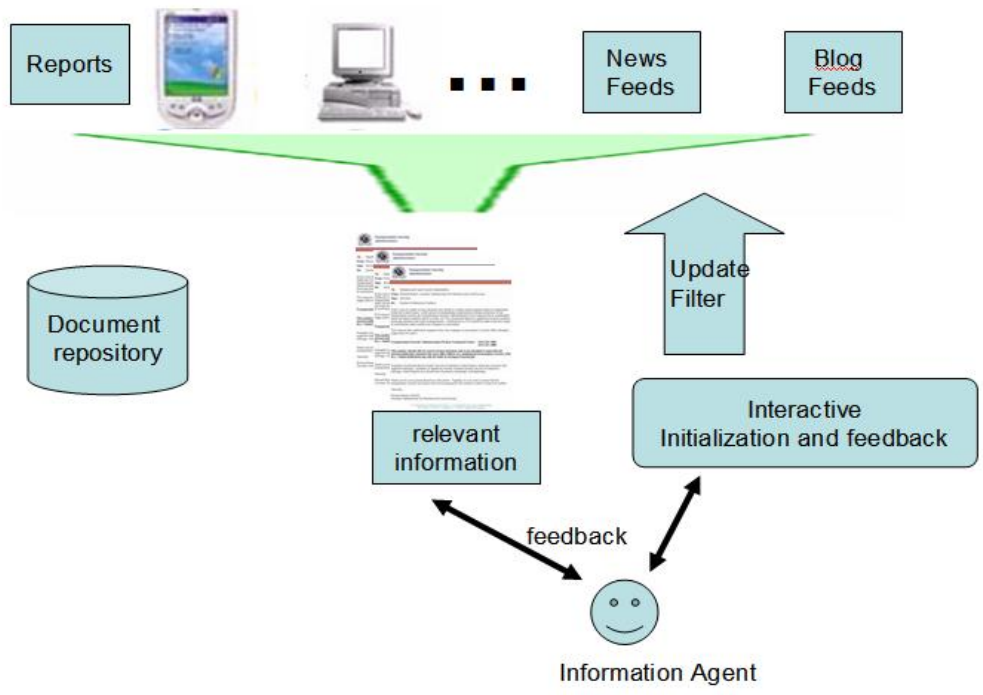

Figure 1: Architecture for a Adaptive Filtering System used by an information agent for tracking information related to potential terrorism.

PROJECT REPORT

Problem 1: To help avoid unanticipated security events such as 911 from happening, agents working on homeland security need to read a large amount of information from heterogeneous information resources (reports from other agents, email, news pages, blog space, internet discussion groups etc.). Unfortunately, it is impossible to read all documents from massive incoming data streams. The nation is at great risk due to the biases of individual agents and misses of information related to potential terrorism attacks.

Problem 2: Since 2005, more than 54 millions Americans were affected because of reported information leaking cases (Labs, 2006). This number only includes publicly reported case and possibly worse case where an organization like FBI does not even know whether or what sensitive information were leaked out (New_York_Times, 2007)(Wasington_Post, 2007)(USA_Today, 2007)(Associated_Press, 2007).

Solution: Adaptive information filtering is a solution to solve the above problems. An **information filtering** system monitors document streams to find the documents that match a user profile. As the system runs, an **adaptive** filtering system frequently updates the user profile based on observations of the document streams, explicit and implicit user feedback.

Figure 1 is a rough architecture of the filtering system. When it is initially launched for a new user, the user uses the interactive feedback interface to provide relevance feedback by identifying examples of relevant and non relevant information. From the training data, the system learns a profile. The profile is used to monitor incoming document streams and identify new relevant information, which will be put into a Relevant Information queue for the user to read.

Research Contribution: This project solved two major challenges in an adaptive filtering system: First, how to build an initial user profile with minimal user effort. Second, how to improve

personalized recommendation based on multiple evidences, such as social networks, implicit user feedback, and explicit user feedback and context information. More specifically, the contributions are:

1. We studied how to develop complex data driven user models that go beyond the bag of words model using the graphical modeling approach (Zhang, 2008). Experiments on data collected from a news filtering user study, as well as another data set, demonstrated that the graphical modeling approach helps us to better understand the complex domain and improve the adaptive information filtering performance.

2. We developed a Bayesian hierarchical modeling approach, which we call Discriminative Factored Prior Models (DFPM), for information filtering (Zhang & Zhang, 2010a). Compared with existing approaches, this approach can 1) borrow discriminative criteria of other users while learning a particular user profile through a factored prior; 2) trade off well between diversity and commonality among users; and 3) handle the challenging classification situation where each class contains multiple concepts. Experimental results on digg.com users show that our models significantly outperform the baseline models of L-2 regularized logistic regression and the standard Bayesian hierarchical logistic regression models.

3. We also studied how to do recommendation with networked data from Internet, social Networks, scientific paper citations, etc (Chi *et al.*, 2009). We first identify four main data dimensions that are common in most of networked data, namely people, relation, content, and time. We propose a polyadic factorization approach to directly model all the dimensions simultaneously and an efficient implementation of the algorithm that takes advantage of the sparseness of data and has time complexity linear in the number of data records in a dataset. Applying the technique on blogosphere and personalized recommendation in paper citations, we found that the framework is able to provide deep insights jointed obtained from various dimensions of networked data.

4. We participate in TREC 2009 relevance feedback track (Zhang *et al.*, 2009). We try clustering and Transductive Experimental Design (TED) methods to automatically find good documents for user to provide relevance feedback. We do query expansion based on a relevance language model learnt on the labeled relevant documents. Our retrieval results after relevance feedback is ranked 2nd among all participants.

5. Standard information retrieval models usually focus on relevancy, without considering other criteria (cost, readability, novelty and recency etc.). We research multi-criteria decision analysis for information retrieval. We found using multiple user-centric criteria always produced better results than a single criterion, and we also found non linear interaction among criteria (Wolfe & Zhang, 2009)(Wolf & Zhang, 2010).

6. Motivated by the commonly used faceted search interface in e-commerce, we investigated interactive user profile learning mechanisms for personalized filtering and personalized retrieval based on faceted document metadata (Zhang & Zhang, 2010b). Experiments on user feedback collected through Amazon Mechanical Turk show that the widely used Boolean filtering approach doesn't work well for text document retrieval, due to the incompleteness of metadata assignment in semi-structured text documents. Instead, a soft model we proposed performs more effectively for personalized retrieval.

7. Since there is no good teaching/tutorial material on adaptive information filtering, we wrote a book chapter on it (Zhang, n.d.).

This project results a demo filtering system that can filter many RSS feeds (including twitter feeds) to prevent critical information from being ignored, which is a serious problem for many companies, government agencies and individuals. The research results have be disseminated widely through high-quality academic journals (eg. Journal of Information Processing and Management, IEEE Transactions on Multimedia) and international conferences (eg SIGIR, CIKM, and TREC).

References

Associated_Press. 2007 (Feb 13). *1.8 million altered after VA data loss.*

Chi, Yun, Zhu, Shenghuo, Hino, Koji, Gong, Yihong, & Zhang, Yi. 2009. iOLAP: A Framework for Analyzing the Internet, Social Networks, and Other Networked Data. *IEEE Transactions on Multimedia.*

Labs, Percept Technology. 2006. *Information Leak Prevention Accuracy and Security Test.* Tech. rept.

New_York_Times. 2007 (Feb 13). *F.B.I Lags in Securing Its laptops and Weapons.*

USA_Today. 2007 (Feb 13). *Audit:FBI has lost more latpops, guns.* USA Today.

Wasington_Post. 2007 (Feb 13). *FBI Reports On Missing Laptops and Weapons.* The Washington Post.

Wolf, Shawn, & Zhang, Yi. 2010. Interaction and Personalization of Criteria in Recommender Systems. *In: Proceedings of User Modeling, User Modeling, Adaptation and Personalization Conference.*

Wolfe, Shawn, & Zhang, Yi. 2009. User-centric multi-criteria information retrieval. *In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.*

Zhang, Lanbo, & Zhang, Yi. 2010a. Discriminative Factored Prior Model for Personalized Content-Based Recommendation. *In: Proceedings of ACM CIKM Conference on Information and Knowledge Management.*

Zhang, Lanbo, & Zhang, Yi. 2010b. Interactive User Profile Initialization based on Faceted Feedback. *In: Proceedings of the 33rd ACM SIGIR Conference.*

Zhang, Lanbo, Zhang, Yi, de Arma, Jadiel, & Yu, Kai. 2009. UCSC at Relevance Feedback Track. *In: Proceedings of the 17th Text REtrieval Conference (TREC), National Institute of Standard Technologies.*

Zhang, Yi. *Text Mining: Theory, Application, and Visualization.* Chapman & HallCrc Data Mining and Knowledge Discovery Series. Chap. Adaptive Information Filtering.

Zhang, Yi. 2008. Complex Adaptive Filtering User Profile Using Graphical Models. *Journal of Information Processing and Management*, 1886–1900.